PROJECT REPORT

IE 590 - ADVANCED DATA ANALYTICS

Statistical Modeling and Analysis Results for Life Insurance Data (Prudential Life Insurance Company Assessment Data- Kaggle)

Submitted to:

Prof. Roshanak Nateghi

Department of Industrial Engineering

Purdue University

Report Prepared by:

Sripati Kannan

Graduate Student

Industrial Engineering

Purdue University

29th April, 2016

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report summarizes the statistical modeling and analysis results associated with the Life Insurance Assessment data released by Prudential Life Insurance on Kaggle. The purpose of this report is to highlight the sampling methods used and the corresponding statistical modeling and inference techniques used during subsequent analysis.

The first part of the report includes the introduction and the background of the report. It includes the problem statement and the goal of the project. The report later highlights the description of the dataset which includes the variables provided and the nature of the response and dependent variables. This will be useful in the exploratory analysis and initial dataset visualization which is really important in any statistical modeling problems.

The second part of the report will have the literature review which includes the approaches followed by some of the existing research done on such data problems and how we could either use them or introduce unique new methods of approach to similar dataset to get the most of the dataset and better predict the response variable.

The third part highlights the importance of sampling the data for proper results for cross validated error which will help us improve the model for future data. It will also contain various summary statistics which can help us understand the dataset properly.

The fourth part will highlight the data analysis method and models used to find the relationship between the response variable and other variables. In-depth explanation of models and to identify which model gives us the best prediction on the test data. Selecting models based on best fit, less bias and variance and the best prediction results using mean squared and mean absolute error of prediction.

The final part of the report will include conclusion, references and appendix which will contain technicalities which was not considered important to be included in the report.

The report reveals that Random Forest is a good model fit for the dataset provided due to its flexibility to handle large dataset and ensemble learning methods. It is a good estimator for categorical variables. Although it is hard to interpret, we can still use it as our aim is to predict and not interpret. Complex models are acceptable as long as the prediction is better. Bias Variance tradeoff needs to be taken into account for complex models as they can tend to overfit the data leading to better results to only one set of data.

INTRODUCTION

Life Insurance application process is still old-fashioned requesting extensive information and medical history records from the customer to analyze risk classification and eligibility. The medical test as a part of the application process is an intensive procedure which is time consuming. The whole process takes as long as 30 days to complete. Life insurance application is known to be detailed asking too much about the client and also requesting a medical test to most of the clients which is used to check the information provided. Insurance premium and the risk associated with a client.

Problem Statement:

People are discouraged to take life insurance products as it is a tedious and a time consuming product. Only 44% of U.S. households own individual life insurance. We need to make this whole process quicker and less labor intensive so that the customers can get a quote quickly while maintaining privacy boundaries.

Goal of the Project:

To develop a predictive model that can quickly classify the risk of selling the insurance product to a customer by analyzing the various variables of data extracted from the client. This model will help companies better understand the power of data points in the existing assessment which can help speed up the process effectively.

Approach for this Project:

Conduct initial analysis to find out which variables impact the response variable and using it in a statistical model with supervised learning methods to find out the impact and predict the risk of giving the insurance product for the applicant.

Developing models which can correctly estimate the risk associated with an applicant is the aim and we will use supervised learning methods for this purpose as the Response variable and corresponding dataset is available for us to help estimate the risk better.

The dataset involves a lot of missing data for important variables as it may be due to lack of disclosure from the applicant which makes it harder to include in the model. However, we can estimate the data using many other tools like MICE package which will be explained in detailed in this report. Such tools provide us better data points to make accurate prediction as raw data obtained needs to be transformed before using it properly.

LITERATURE REVIEW

Wikipedia defines “Predictive Risk analysis” as using historical and current data to predict the risk of a certain activity. This is done using various statistical models measures dependency of response variable among many other factors to allow assessment of risk associated with a particular set of conditions, guiding decision making for different events.

Deloitte Consulting LLP (2010) published a research paper which talks about the various aspects of risk analysis with respect to cost and historical records of an applicant which will help them predict the effectiveness of a product. Brockett, Xia and Derrig (1998) apply a feature mapping process to classify potential fraud cases in bodily injury claims. Tennyson and Salsas-Form (2002) look at a small sample of claims to determine if auditing was used primarily for detection of fraud or deterrence of future fraud.

Life Insurance Risk Analysis and Costing (2008), a research paper published by a task force talks about the impact of various factors like product variations, term and premium rate. This paper discusses a lot of factors which affect the risk of an applicant and the various costing factors. It highlights various risk elements associated with an applicant. We can use this information to weigh different aspects of risk and analyze the same for out dataset.

Most of the research done focuses on risk of cost reduction and product impact. Some of the papers talks about the impact of the applicant and historical records which will have on the insurance product, although not utilizing it to predict the risk associated with the applicant. This paper focuses on the applicant and the data provided of the applicant along with other factors such as product information, previous insurance information and employment information which can have a direct impact on the response variable. Response variable is the risk associated with an applicant such that he might either die early or he may not be able to pay the premium.

We will find the relationship between the dependent and the response variables with the help of various statistical modeling techniques which will be highlighted in depth in this report. This can help us better understand the scenarios in which risk is associated with the applicant and also what variables will be required to predict the risk of the future applicant. We can reduce the input of data which can improve the lead time associated with the application of each client eventually getting more clients to sign up for the company’s insurance.

INFORMATION ON DATA:

Data Source:

This dataset is released by Prudential Life Insurance as a competition on the website of Kaggle. The link is as follows: *https://www.kaggle.com/c/prudential-life-insurance-assessment*

Dataset Description:

Variables: This dataset has 127 variables which describes various attributes such as physical measurements and medical history which was provided with the application for the insurance. It also has a response variable which is the target variable relating to the final decision associated with the application. It has 8 levels which conveys the extent of risk associated with selling the insurance to the customer. Higher response variable indicates higher risk.

Variable Types: The dataset has different variable types like nominal, continuous and discrete variables. There are a few dummy variables which has inconsistent impact on the data but which also needs to be taken into account as they impact the decision variables irregularly.

Distribution: The data set is divided into two major sets: Training and Testing Dataset. The Training dataset has 59381 observations with 128 variables including the response variables. This data set will be used to train and test the model to improve its predictability and dependency of other variables on the response variable. Test data set will be bootstrapped from this dataset so we can train the model on training dataset and predict with the test dataset.

Model Implementation: A model will be developed based on the behavior and associations of variables with the response function. The model with the highest accuracy will be applied on the test data to help make the application process faster and efficient. This will also help us find the interactions of various factors on the risk of selling insurance to the customer.

The dataset includes variables that are categorical, ordinal and continuous in nature. The list of all the variables and the

Data Summary:

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of applicant
Ht	Normalized height of applicant
Wt	Normalized weight of applicant
BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

The following variables are all **categorical (nominal)**:

Product_Info_1	Employment_Info_2	InsuredInfo_4	Insurance_History_3	Medical_History_29
Product_Info_2	Employment_Info_3	InsuredInfo_5	Insurance_History_4	Medical_History_30
Product_Info_3	Employment_Info_5	InsuredInfo_6	Insurance_History_7	Medical_History_31
Product_Info_5	InsuredInfo_1	InsuredInfo_7	Insurance_History_8	Medical_History_33
Product_Info_6	InsuredInfo_2	Insurance_History_1	Insurance_History_9	Medical_History_34
Product_Info_7	InsuredInfo_3	Insurance_History_2	Family_Hist_1	Medical_History_35
Medical_History_2	Medical_History_8	Medical_History_16	Medical_History_22	Medical_History_36
Medical_History_3	Medical_History_9	Medical_History_17	Medical_History_23	Medical_History_37
Medical_History_4	Medical_History_11	Medical_History_18	Medical_History_25	Medical_History_38
Medical_History_5	Medical_History_12	Medical_History_19	Medical_History_26	Medical_History_39
Medical_History_6	Medical_History_13	Medical_History_20	Medical_History_27	Medical_History_40
Medical_History_7	Medical_History_14	Medical_History_21	Medical_History_28	Medical_History_41

The following variables are **continuous:**

Product_Info_4	Employment_Info_6
Ins_Age	Insurance_History_5
Ht	Family_Hist_2
Wt	Family_Hist_3
BMI	Family_Hist_4
Employment_Info_1	Family_Hist_5
Employment_Info_4	

The following variables are **discrete:**

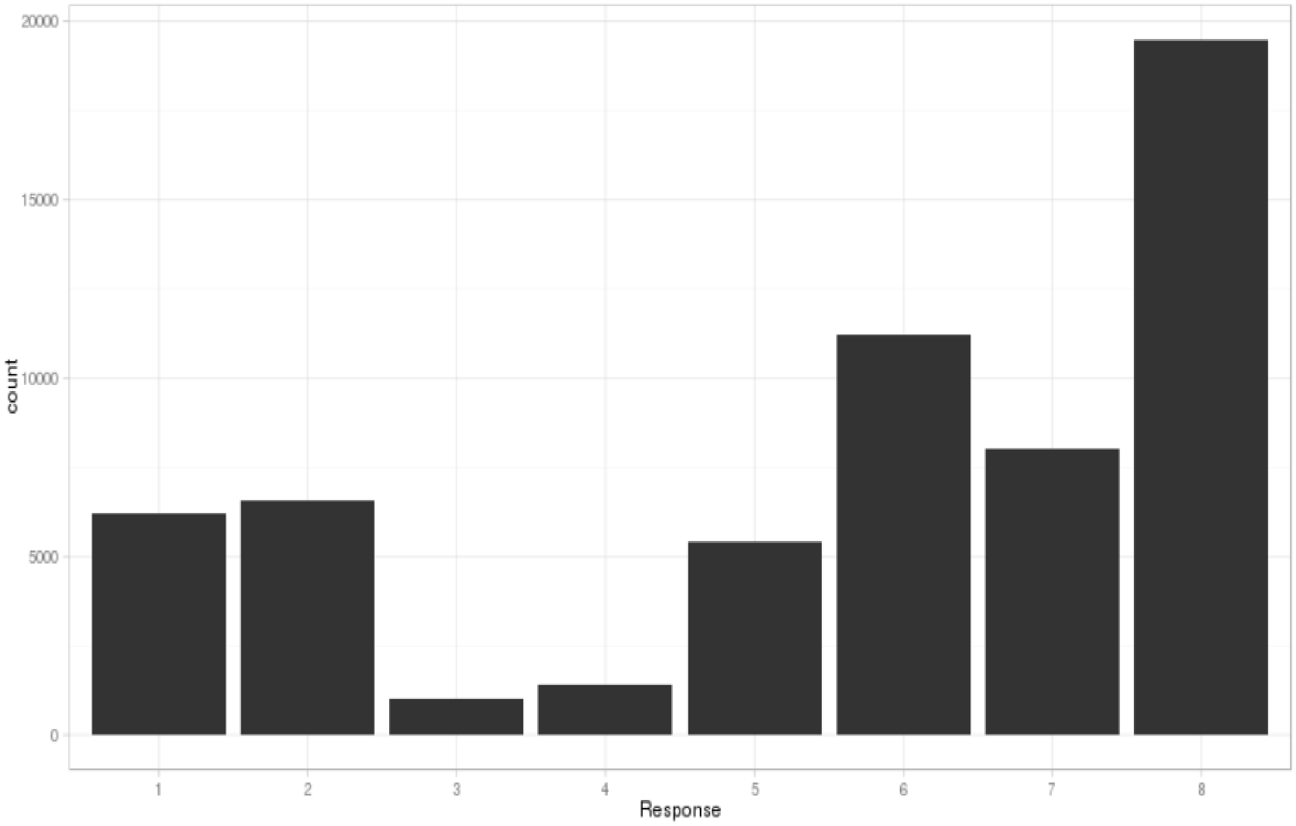
Medical_History_1	Medical_History_24
Medical_History_10	Medical_History_32
Medical_History_15	

Medical_Keyword_1-48 are dummy variables.

SAMPLING METHODS

As the dataset is large, we will need to use various sampling techniques to get the best results out of this dataset. The Sampling technique used for models in this analysis is Random Sampling method wherein the samples from the data set are taken randomly and then part of it is used for training and part of it is used for prediction. This technique is called as Cross Validation. K-Fold Cross validation is used for large dataset. The dataset is divided into k parts out of which one part are hold out for prediction and the model is run for the other k-1 parts to find out the prediction. I used k= 4 with iterations 5-10 depending on the model used. As random forest's computation time is high, we use less iterations and with simpler models, I used 10 iterations. We can calculate Mean Squared and Mean Absolute Error respectively for a no of time on the model to check predictability and error in test and training data set sample. It also measures the model's prediction as a value when comparing different models.

EXPLORATORY DATA ANALYSIS



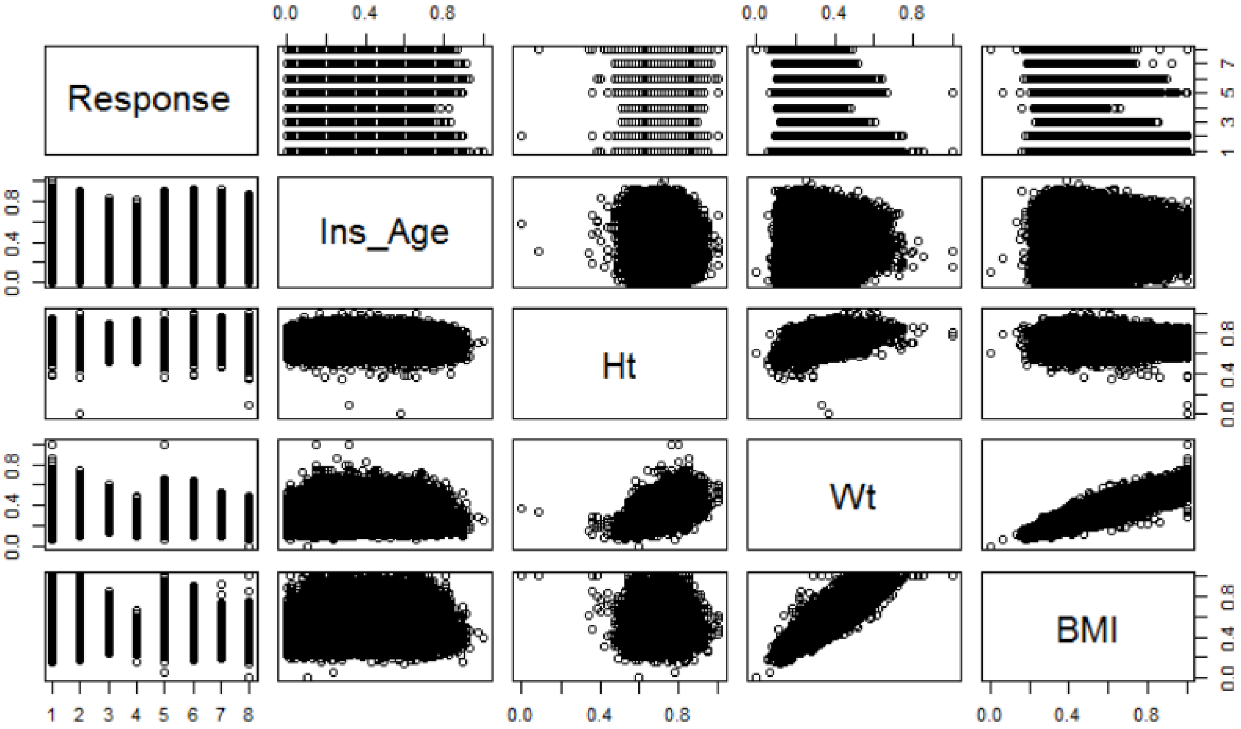
Histogram of Response Variable of Training Dataset (Values range from 1-8)

We can see that the response variables have more than three quarters of the data over 5 and above which tells us the risk associated with the customers in the training data set.

We can develop a model targeting the variables which have a larger impact on the response variable and trying to reduce and remove the variables which have negligible impact on the response variable. With this, we can reduce the time consumed in predicting the risk associated and improve efficiency of the application process. This will result in taking less information from the applicant as a result of which the efficiency of this process can increase. This is the objective of this study which will be achieved with model development and the variable reduction using different approaches learnt in class.

Correlation Matrix Plots:

This plot shows us the pairs plot of matrix of some important variables which directly impact the variable.



These plots help us understand the relationship between response and other variables. We can see that BMI and weight are linearly related as they are a function of one other. Response variable is a categorical variable and hence we can see the respective relationship between the other variables. These plots provide insight into relationship between other variables. Some more plots have been added to the appendix for reference.

Sign Rank Test Results

Sign rank test results are used to evaluate the best model using the MSE and MAE values to find out which one is statistically significant. It is compared with two objects of values obtained from the model respectively. The null hypothesis of the test that the means are equal must be reject and we should check for the estimate of values. While Wil-Coxon test can be used for non-normal data, similarly welch can be used for normal data. We assume and the visualization data show that the data is non-normal.

Paired t-Test

It is used in case to check whether the population means in both case is correlated. The null hypothesis is that they are same. We reject or accept based on the test results which is indicated by the respective t or p values for the parameter.

INTIAL MODEL DEVELOPMENT AND THEIR INFERENCES

We are to compare the behavior of models on test dataset by training it using the train dataset. The model is better predicted using cross validation with k as 4 because the dataset is quite large and we need to make sure the variance is explored in all the holdouts. We find the Mean Square Error and Mean Absolute Error for each model using cross validation. Also, we check the accuracy of the model with the test dataset to check for the fitness of the model.

In this dataset used, the variables with the missing values are ignored as most of the data for those variables are unavailable which maybe harder to approximate as to which variables will have a significant impact on the response variable

The models used are as follows:

<u>Model 1: General Linear Model</u>

General Linear model is used to understand and develop a simple model which can help us explain a linear relationship between the response variable and other dependent variables.

Some of the assumptions and benefits of the linear model include:

1) Linear Relationship
2) No auto-correlation
3) Constant Variance
4) Simple model and easy to interpret.
5) No or little multi-collinearity
6) Can handle categorical values

The equation of general linear model is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i.$$

General Linear model provided a R square value of 22.16% which is poor and does not show many variations in the model. It has a good prediction power but the R square is less which means all the variance is not explained by the variables included in the model.

<u>Model 2: General Additive Model (GAM)</u>

GAM provide a model which can help us use splines to account for variability in the data. It is more useful than general linear model as it helps us not to restrict ourselves to the linear relationship, but to use different splines for different variable to account for variance. However, their prediction power is not satisfying in this case.

Some Assumptions and positives of GAM are:

1) Different smoother functions can be used for different variables to explain variability.
2) Fast algorithm which can be used for prediction as well
3) Explains nonlinear relationships

GAM induces issues related to prediction. Every variable may have different curves which needs to be associated correctly to the spline function. Little complex to develop the model. However, its prediction in this case was not that great compared to Random Forest and BART.

The equation is as follows:

$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m).$$

The functions in the right side of the equations are splines functions which are associated with the dependent variables to develop a better model.

Model 3: Multivariate Adaptive Regressive Splines (MARS)

MARS model is an improvement to the general linear model designed to automatically accommodate non-linear relationships in the dataset.

The equation is as follows:

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

This equation represents the basis function which includes the constant and the hinge function which is a weighted sum of all the variables in the dataset needed to predict the response variable.

MARS provide the one of the best MSE and MAE values as well as provide the best accuracy for the test data set. However, this too gives a reasonably less prediction power.

Model 4: Classification and Regression Trees (CART)

CART model is a tree based model which helps classification and regression problems. It makes a decision trees based on different start and end nodes. Bagging and boosting operations will be performed on the CART to make it robust and to improve prediction power. The CART model can overfit and the tree length needs to be decided properly. Hence, Pruning will help reduce the model in which trees are made so that it can predict better.

CART model can be easier to interpret but it can be complex sometimes due to many features present in the training dataset which means more trees. Without pruning, many trees will be formed without any proper cut-off nodes resulting in increasing complexity of the model.

CART provide decent values of prediction power however, its MSE and MAE values are high. It is not a very good model for this case and also due to many variable interactions, the decision tress becomes complex.

Model 5: Random Forest:

Random Forest is an ensemble learning method designed for classification and regression problems. It uses number of trees as a classifier and gives us the prediction which is robust. It also gives us variable importance plots which can be useful to see which variable has the higher impact on the response variable.

Random Forest is the best model for this dataset as it takes into account all the possible scenarios and comes up with the best classification tree to classify our response variable with the dataset provided. Not much of the variance is explained as the missing variables are not taken into account for model selection.

Model 6: Bayesian Additive Regression Trees (BART)

BART model uses Bayesian conditional probability using a prior and a posterior to develop the model. It has a robust algorithm as the trees are made using conditional probability. BART provided a decent result but Random forest had the best predictive power for out of bag samples.

The equation of sum of trees model is as follows:

$$Y = \left(\sum_{j=1}^{m} g(x; T_j, M_j) \right) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

Where Tj is a binary regression tree and e is the error associated with each term that follows a normal distribution.

BART is a good model to predict out of bag errors, but in this case, it tends to perform a little less than Random Forest.

Model 7: Mean only Model

Mean only model allows us to compare our variables with mean so that we can see if what the variance of the variable is in relation and if it is useful in explaining the variability in the model. It is calculated as (training data response-mean(test data Response)). The mean square error and mean absolute error is calculated and inserted in the table below.

Model Used	Mean Square Error	Mean Absolute Error	Standard Deviation	Accuracy (%) for out of bag samples
Linear Model	4.65	1.73	0.017	15.32
Generalized Additive Models	4.423	1.655	0.0315	17.27
MARS	4.15	1.5813	0.005	17.8
CART	4.84	1.786	0.0092	10.10
Random Forest	3.67	1.32	0.00811	25.23
BART	3.86	1.49	0.0032	21.32
Mean Only	5.95	2.01	0.023	-

The table above represents the Mean Square error, Mean absolute error and the accuracy of the models with respect to the out of sample test dataset. Random Forest is the best with the maximum accuracy and the least error terms.

Model Significance

Model	Hypothesis Accepted
GLM and GAM	Null
	(Accept GLM)
GLM and MARS	Alternate
	(Accept MARS)
CART and MARS	Alternate
	(Accept MARS)
RF and MARS	Null
	(Accept RF)
RF and BART	Null
	(Accept RF)

Wilcoxon Paired Sample test was performed on MSE and MAE vectors for different models to test if there are statistically significant or not. It turns out even though, RF, BART and MARS have similar prediction errors, then tend to be significantly different from one other. Hence, we choose Random Forest from the table above as it has the best out of bag sample prediction power due to its robust algorithm.

HANDLING MISSING VALUES

The dataset provided has many missing values for certain important variables due to which model developing is little complex. Thirteen of the 128 variables had missing values. These needed to be taken care off as the variance in the model was not explained properly and these variable needed to be added in model. These variables have missing values in random order which needs to be estimate to check their influence on the response variable. The missing values was also accounted for by using is.na function in R for the train dataset.

Two approaches where used to handle missing data in this dataset:

1) Elimination of variables which had missing values more than 50% observation.
2) Using Multivariate Imputation by Chained Equations(MICE) package in R to find out and approximate the missing data

Mean and Median Substitution was not used as it provides a lot of bias in the training dataset which could result in improper prediction of the response variable.

1) Elimination of variables which had missing values more than 50% observation.

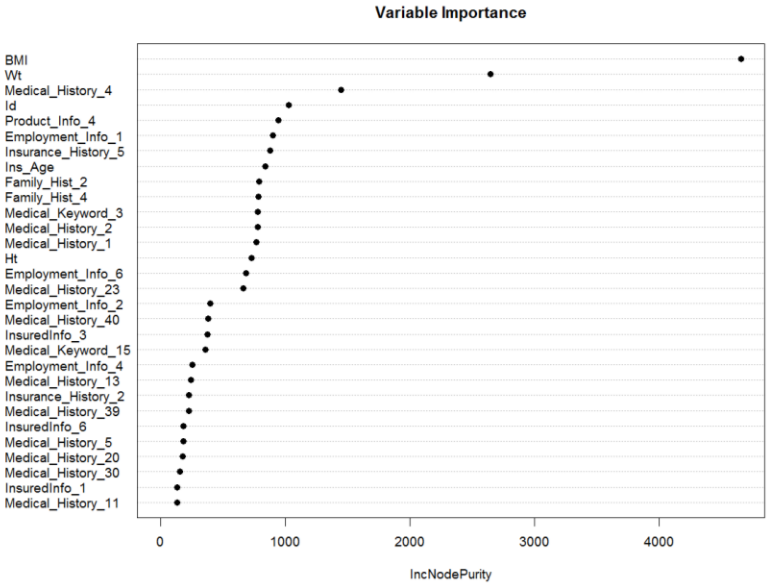
Model Used	Mean Square Error within training dataset	Mean Absolute Error within training dataset	Standard Deviation	Accuracy of predicting test data set (out of bag samples)
General Linear Model	3.28	1.29	0.011	28.07
Generalized Additive Models	3.52	1.39	0.010	22.8
MARS	3.25	1.28	0.0015	28.9
CART	3.67	1.40	0.0062	21.34
Random Forest	3.08	1.26	0.0081	33.2
BART	3.18	1.28	0.0097	30.3
Mean Only	3.8	1.56	0.035	18.21

The dataset which had more than twenty percent of the missing terms in the dataset, those variables were ignored while modeling as we need a fixed number of samples to estimate and higher degree of freedom to calculate the p value and error terms.

Model analysis was performed as earlier with the same models used and cross validation performed for each and the MSE and MAE vectors were found out.

This was done for a full model with all the variables included in the model for all models as the variance explained by the model reduces when variables are reduced which results in poorer results impacting the MSE and MAE vector. Total observations were reduced to 8176 observations and analysis was performed on this dataset.

OPTIMAL MODEL 1: RANDOM FOREST



The selected model is Random Forest as it as the best MSE and MAE vectors which is related to predictability. It explains a decent percentage of the variability of the dataset upto 35% which is the best among the models. Also, the accuracy computed after the Misclassification error reported by Random Forest is 33% which is way better than the other models.

The partial dependency plots for some variables are attached in Appendix.

t-Test for significance:

To check the statistical importance of the model among other models, we do a t test of Wilcoxon for non-normal data to find out the importance. As explained before, we need to check and estimate the mean of MSE and MAE vectors between two models to find out the significance between models.

t- Test:

```
data:  vecMAErf1 and vecMAEcart
t = -12.61, df = 6.6713, p-value = 6.713e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1635521 -0.1114626
sample estimates:
mean of x mean of y
1.265079 1.402586
```

0 is not included in the CI and hence the model vectors are significantly distinct. We can see that the estimate is less for Random forest. Hence, we choose Random forest which is consistent with the other deductions.

2) **Using Multivariate Imputation by Chained Equations (MICE) package in R to find out and approximate the missing data**

Missing data points are very tedious to deal with. We can either ignore the values that are missing and use data points that are available or we can use mean substitution to deal with it. But these may cause bias in the dataset.

The MICE package in R, helps you imputing missing values with possible data values. These possible values are drawn from a distribution specifically designed for each missing data point.

Using m=5 iterations for MICE package, we estimate the missing values for each data point that is missing. This will help us use the variable which has less data pints without inducting bias and will help us develop a good model which may provide good prediction results.

This is a good strategy and gave optimal results with less MSE and MAE error vectors.

FINAL MODEL: RANDOM FOREST WITH MICE PACKAGE

We know that random forest is providing better results than the other models as it has good Rsq, MSE and MAE vectors and is statistically significant.

We can develop a model with MICE and random forest which should provide optimal results. MICE will help approximate the missing data points and we might be able to use variable which were ignored earlier due to missing data points.

From the earlier Random Forest and the repeated runs of this model, we can find the variable importance of each variables with the response variable. Hence, we can use the variables which have a variable importance of more than 200 as a general number with an assumption. This is to reduce the complexity of the model and to find the tradeoff between using all the variables, explaining the variance and also find the best prediction.

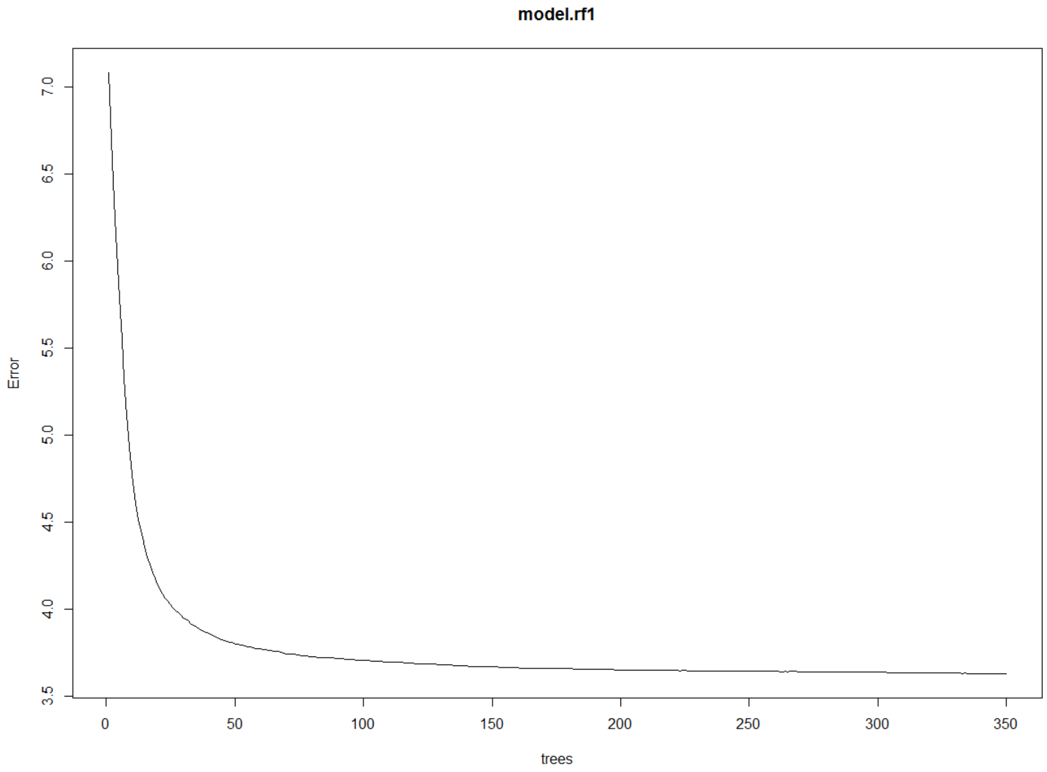
Model Description:

30 Variables used in the model which have a higher impact on the response variable. A random forest model is run with the number of trees first set at 500 default. It was reduced to 350 in the subsequent runs as the curve became stagnant after 350 trees. This enables to produce a compact model. Many variables were converted to factors as they tend to provide a good result in Random forest as it helps classification better.

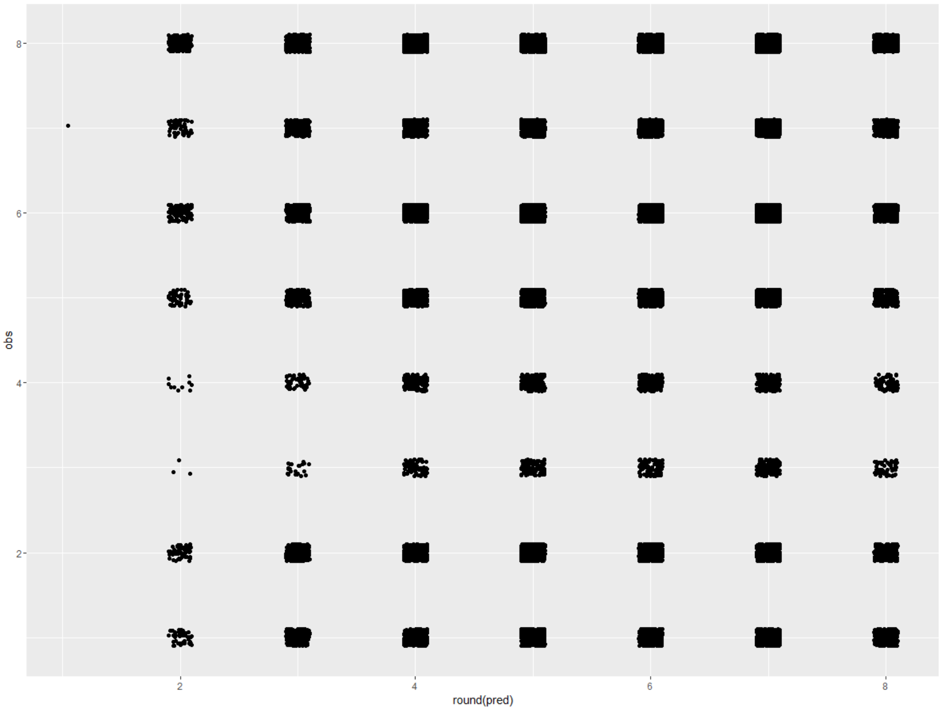
Cross validation with k=4 was performed for this large dataset as it is a proper parameter that estimates the error and the optimal one. Number of iterations was at 5 as the computation time for Random Forest is very high. It takes 2 hours to build one model in one iteration.

MODEL DIAGNOSTICS

Let’s look at some diagnostic plots for Random Forest.



This plot shows us that the above model can be optimized with number of trees upto 150 as there is not a lot of change in the error for increase in number of trees. More trees result in overfitting of the data which can results in bias when the out of bag samples are considered. Hence, we can optimize the parameter and set it at 150 to minimize the prediction error.



The above plot shows the observed versus the predicted values of the test dataset. As you can see, the accuracy of prediction is not that great. It can be associated with lots of missing values and outliers in the dataset which makes a bad model.

The dataset also has a lot of ordinal values which are of different levels having different impact on the dependent variable. It is harder to incorporate the variables accurately predicting the

This was one of the least MSE and MAE error reported in this analysis. MSE= 1.97 and MAE= 1.007. The variance explained was 45% which was the highest.

The accuracy reported after computing the misclassification error is 35% percent which is really the best possible accuracy achieved considering the data set and the outliers. We may still need a better model to explain the dataset and perform and predict better.

We have to perform a t-test to check the significance for the model with other models with the same dataset.

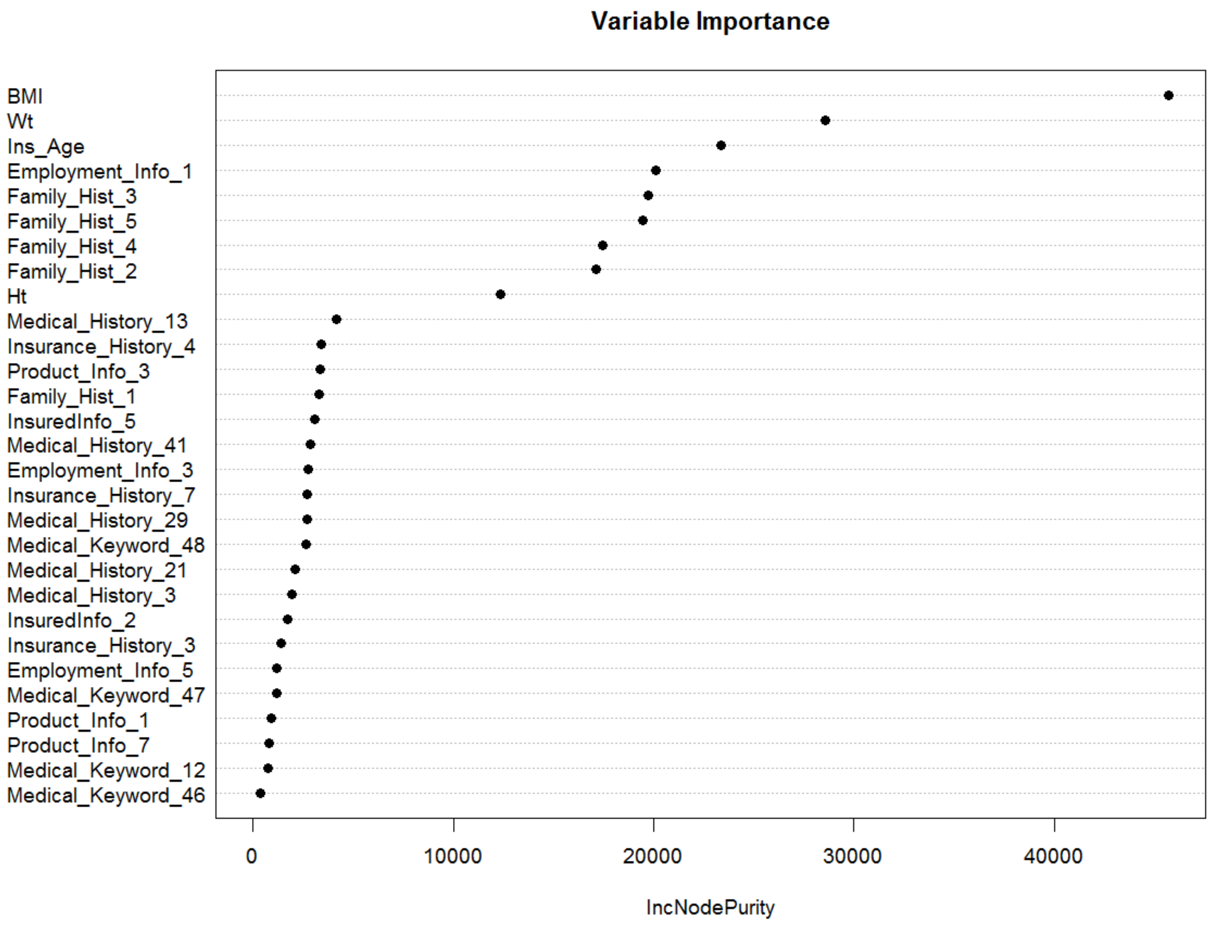
Welch Two Sample t-test

```
data:  vecMSErf and vecMSEmars
t = -0.84557, df = 4.1193, p-value = 0.3441
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5779401  0.3057352
```

sample estimates:
mean of x mean of y
 1.9652 2.546

Accept Alternate Hypothesis. Difference is significant so we use Random Forest as the MSE vec tor is less.

Variable Importance Plot for Random Forest



According to the variable importance plot, it is quite intuitive that BMI and Weight are important when considering risk factor for insurance. Also, the above plots include Employment History, F amily History, Height and Insurance History and so on.

Inference for the model:

The t-test tells us that the test is significant and the means are not equal. We choose Random Forest as it has a lesser estimate of error according to t-test. All the deductions show that the Random Forest is the optimal model for this dataset with MSE, MAE and accuracy on the test data.

CONCLUSION

From the analysis described above, we come to a conclusion that we can predict the risk of an applicant to an extent which is restricted due to too much outliers which was suggested by the data visualization. Much of the variance remains unexplained because of these data points.

The random forest model with 30 variables out of the 128 variables is one of the best models obtained in the analysis with the MICE package. However, this model is not the optimal one as the prediction error is still a little high and not much of the variance is explained.

We will need a better model which can explain much more variance in the data set as well as give a better prediction. Random forest does a good job in prediction however, it's computation time and complexity may be an issue. We need to consider may other tools for analysis in this dataset which will be applied at a later stage.

Limitations

There were quite a few limitations to my analysis. Most of it was due to the dataset selected which had very few information of the variables which made is harder to draw inference from the model selected. The dataset used had a lots of missing values for important variables and had quite a few dummy variables which had effect only on a few readings. It made the analysis harder. Also, the variance was quite high due to the outliers in the dataset which the model could not explain properly. The model selected was Random Forest which has a good prediction power but it could not explain the variance when out of bag samples are considered.

We will need a powerful model with dataset having all the reading present to develop a good model which can clearly the predict the risk of Insurance applicant without bias or error.

Future Work

The models suggested above are not quite accurate enough. We will need to develop better models by using higher order models such as Support Vector Machines, Neural Nets and many more. These models can help us better explain the variance than the models suggested above. These models can also handle missing values appropriately so we can get a more powerful result with them.

As my future study, I plan to incorporate the following:

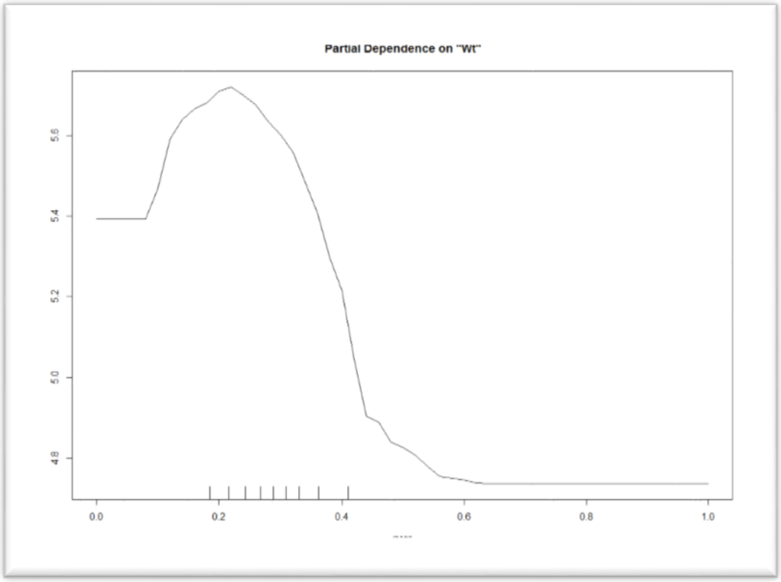
- Improving the prediction of models using methods such as SVM and Neural Net.
- Reducing input data by making a consolidated model which can predict better.
- Faster algorithms with less computation time which can make the application process for insurance faster.

REFERENCES

1) Prudential Life Insurance Competition Dataset- Kaggle
   https://www.kaggle.com/c/prudential-life-insurance-assessment
2) Missing Data plot to show missing value in training dataset
   https://www.kaggle.com/wittmaan/prudential-life-insurance-assessment/exploring-the-data/comments
3) Model Description of GLM, GAM, RF and MARS- Wikipedia
4) Chipman, Hugh A.; George, Edward I.; McCulloch, Robert E. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4 (2010), no. 1, 266--298. doi:10.1214/09-AOAS285. http://projecteuclid.org/euclid.aoas/1273584455.
5) Predictive Modeling in Automobile Insurance: A Preliminary Analysis by Stephen P. D’Arcy
   https://business.illinois.edu/ormir/Predictive%20Modeling%20in%20Automobile%20Insurance%207-1-05%28PDF%29.pdf
6) Research Paper – Life Insurance Costing and Risk Analysis
   Canadian Institute of Actuaries June 2008
7) Risk Management and the Rating Process for Insurance Companies
   http://www3.ambest.com/ambv/ratingmethodology/OpenPDF.aspx?rc=197707
8) Mortality Risk Insurance and Value of Life - Darius Lakdawalla and NBER Julian
9) Imputing Missing data with R, MICE Package
   http://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

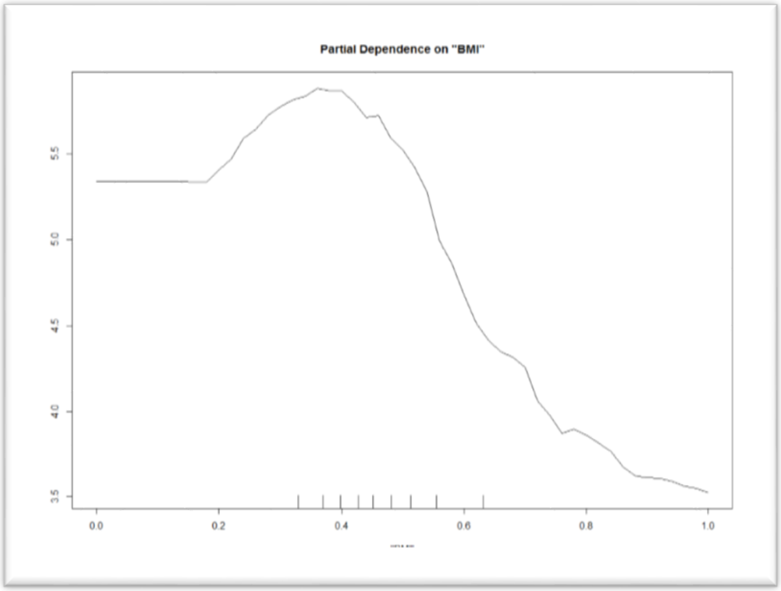
APPENDIX

A) Partial Dependency plot of Response vs Weight



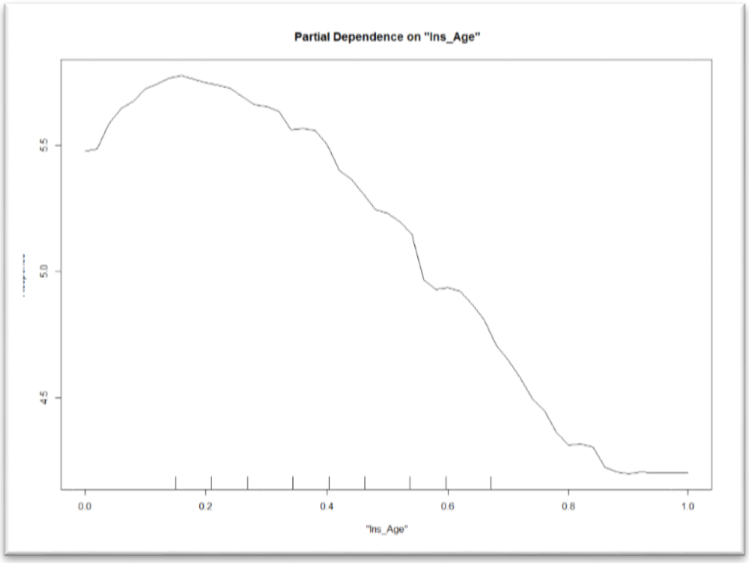
This plot shows that as the weight increases the risk increases. For some weight, risk is high and for very less weight, the risk could be low as well.

B) Partial Dependency plot of Response vs BMI



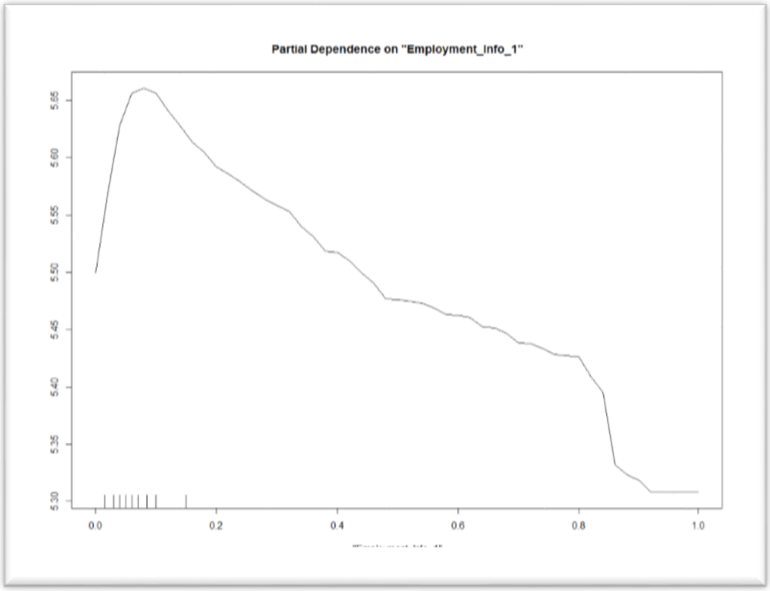
This plot shows that as the BMI increases the risk increases. The risk is constant for some BMI after which there is a slight decrease in risk as for some BMI, it can be safe.

C) Partial Dependency plot of Response vs Age



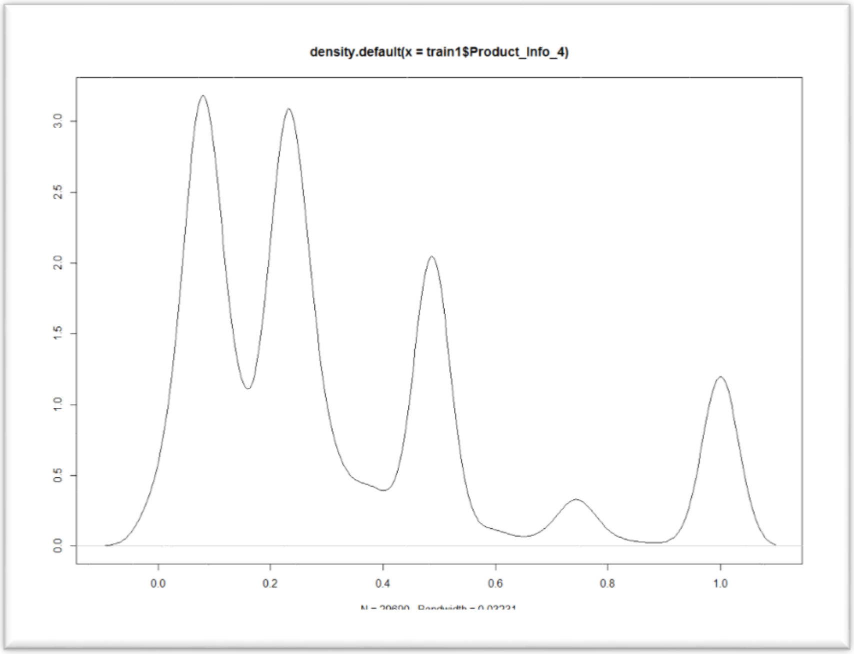
This plot shows that as the age increases the risk increases. There is a peak at the start which indicates, even when the age is very less, there could be some risk involved as children can be risky.

D) Partial Dependency plot of Response vs Employment Info_3



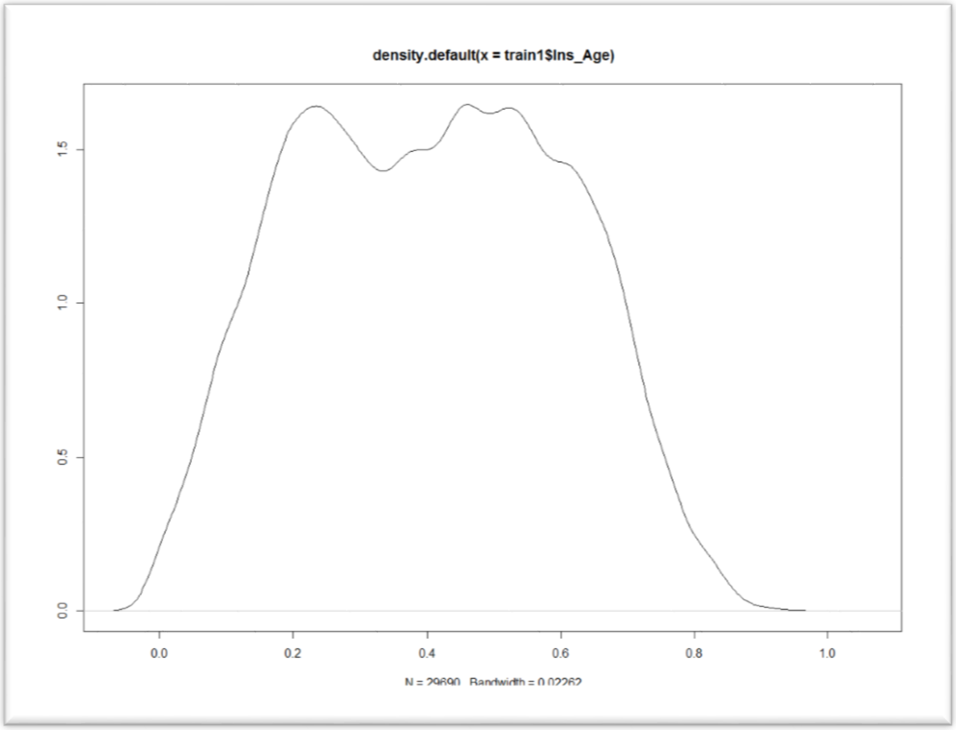
This plot shows that as this variable of employment increases the risk increases.

E) Density Plot for Product Info 4



Shows a wavy pattern with high peaks at intervals

F) Density Plot for Age



Age is Majorly distributed between 0.2-0.7. We can understand the age of the dataset

R-Script for MICE Package:

```
install.packages("mice",repos="http://cran.rstudio.com/",lib="/home/kannans/Rlibs")
library(mice,lib.loc = "/home/kannans/Rlibs")
full <- rbind( train[,-ncol( train )], test )
remove( train ); remove( test )
names <- names( full )

install.packages("foreign",repos="http://cran.rstudio.com/",lib="/home/kannans/Rlibs")
library(foreign,lib.loc = "/home/kannans/Rlibs")
train_imp <- mice( full, m = 1 )

full <- complete( train_imp )
remove( train_imp )
names( full ) <- names
train <- full[1:length( Response ),]
train <- cbind( train, Response )
full <- full[(1+length( Response )):nrow( full ),]
```